Part 2: CLM Practice (Homework 12)

Lab 2: What Makes a Product Successful? - W203 Section 8

Team Herkimer: Rick Chen, Chase Madson, Maria Manna, Jash Sompalli

Jul 25, 2022

Contents

We are given a data set on YouTube videos and want to run the model $ln(\text{views}) = \beta_0 + \beta_1 \text{rate} + \beta_3 \text{length}$, where the variables mean the following:

- `views`: the number of views by YouTube users.
- `rate`: This is the average of the ratings that the video received. You may think of this as a proxy for video quality.
- `length`: the duration of the video in seconds.

The resulting model is summarized as follows:

Table 1: Regressing Log-Views on Rate and Length

	Dependent variable:
	log(views)
rate	0.472*** (0.452, 0.493)
length	0.0005*** (0.0003, 0.001)
Constant	5.411*** (5.324, 5.497)
Observations	9,609
R^2	0.190
F Statistic	1,123.681*** (df = 2; 9606)
Note:	*p<0.1; **p<0.05; ***p<0.01

1 Evaluate the IID Assumption

For this assumption to be met we expect the data to have been drawn from an independent and identically distributed sampling process. The unit of observation of this data set is a distinct YouTube video, and the creators of this sample of YouTube videos describe their sampling process as follows:

> *We consider all the YouTube videos to form a directed graph, where each video is a node in the graph. If a video b is in the related video list (first 20 only) of a video a, then there is a directed edge from a to b .... When processing each video, it checks the list of related videos and adds any new ones to the queue.*[1]

By this description, we know that each video was selected due to its relation to a previously selected video included in the sample. There is an explicit correlation between units of observation and a clear violation of the IID assumption, specifically *independence*.

Moreover, the creators state that they began pulling their sample from an "initial set of videos from the list of 'Recently Featured', 'Most Viewed', 'Top Rated', and 'Most Discussed'...", each one of these likely representing a cluster within the population of YouTube videos. This means the first subset of videos comes from one of these four clusters, and then subsequent data was pulled based on how closely related they were to this first subset. Therefore, we have major presence of clustering in our sample in contrast to the population, another violation of the IID assumption.

When considering whether the observations of this data were drawn from an identical distribution, we cannot easily conclude that the population distribution is the same over time. The method of selecting data heavily relies on the initial set of videos, which comes from the snapshot data of 'Recently Featured', 'Most Viewed', 'Top Rated', and 'Most Discussed' in Feb 22nd 2007. There might be a trend of what people like to see or discuss in a certain period of time and would change in a different time period. Thus, there is a possibility that more than one population distribution presents at different times of selecting data. The "Butterfly effect" of using the initial set of videos will even amplify the data selection biases.

Finally, it is concerning to look at distribution of the `age` variable and see a complete absence of videos between 1 and 325 days old, which is likely a result of the "breadth-first" search algorithm. This tells us that the sampling approach has holes in it and does not resemble a simple random sample.

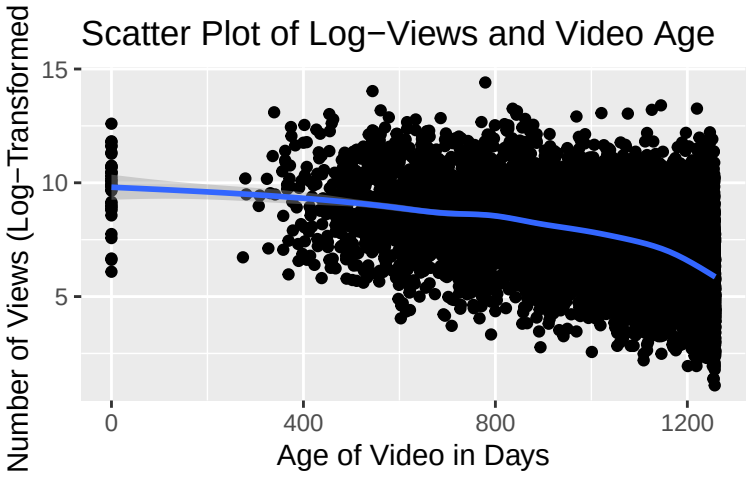

Figure 1: Unexplained Gap in Videos Aged between 1 and 325 Days Old

Conclusion: No, the IID assumption is *not* met.

[1] From *Dataset for "Statistics and Social Network of YouTube Videos"* https://netsg.cs.sfu.ca/youtubedata/

2 Linear Conditional Expectation Assumption

For this assumption to be met we expect to observe no trend (i.e., a horizontal line) when observing the scatter plot between the model's fitted values and its residuals.

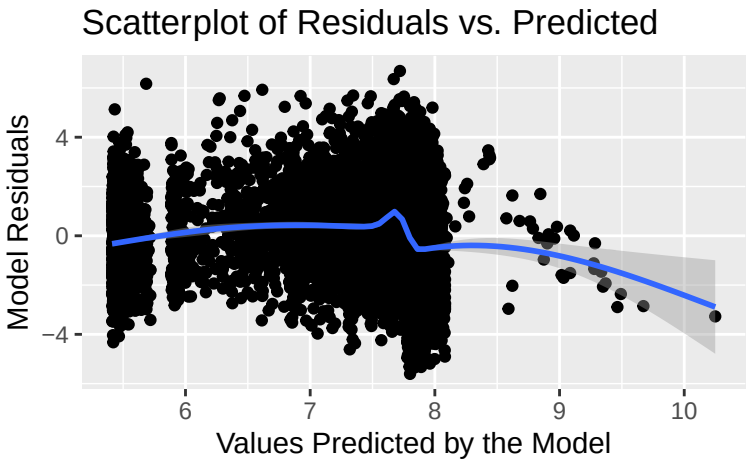

Figure 2: Unusual Trends are Apparent Moving Left to Right

Here instead of a horizontal line we see an unusual trend in the residuals as the predictions increase. This model is systematically overestimating the values at the higher end, with a dramatic rise and dip occurring in the mid-section. This means there is a conditional relationship between residuals and fitted values depending on where we are in the range of predicted values. Therefore, we observe some violation of the linear conditional expectation assumption.

To take a deeper look, let's plot the residuals individually against the two predictors `rate` and `length`.

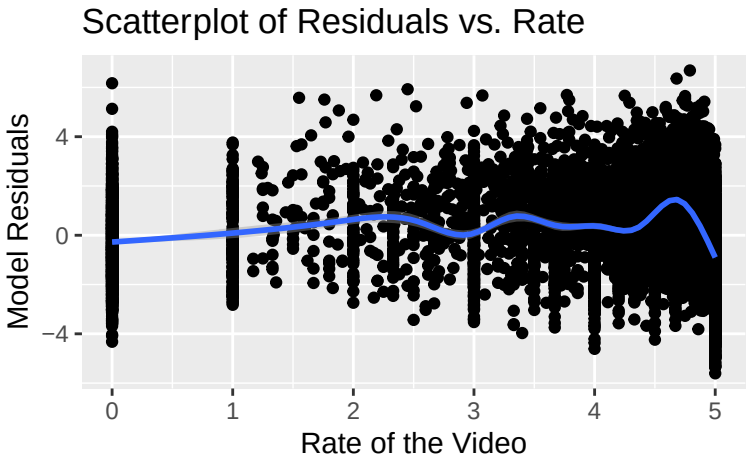

Figure 3: Unusual Oscillations in Residuals as Rate Increases

When plotting the residuals against the predictor `rate`, we see the source of the dramatic rise and dip in the mid-section of the previous plot.

When plotting residuals vs. the predictor `length`, we see the source of the slow downward curve seen in the residual vs. fitted plot. Reasonably speaking, the length of a video likely contributes diminishing returns towards the number of views it receives, for which a linear relationship does not extrapolate well.

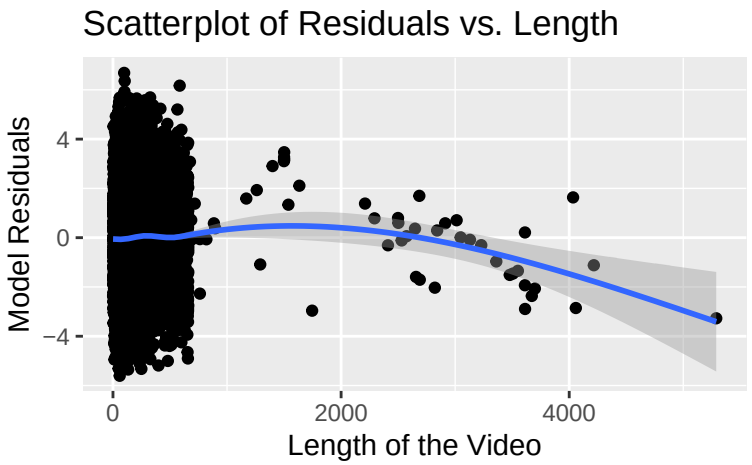

Figure 4: Model Not Accounting for Diminishing Returns in Video Length

Conclusion: No, the assumption is *not* met.

3 No Perfect Colinearity Assumption

For this assumption to be met we expect to find no evidence of a perfect linear relationship between our two predictor variables, `rate` and `length`. To test for this, we plot these two predictors against each other.

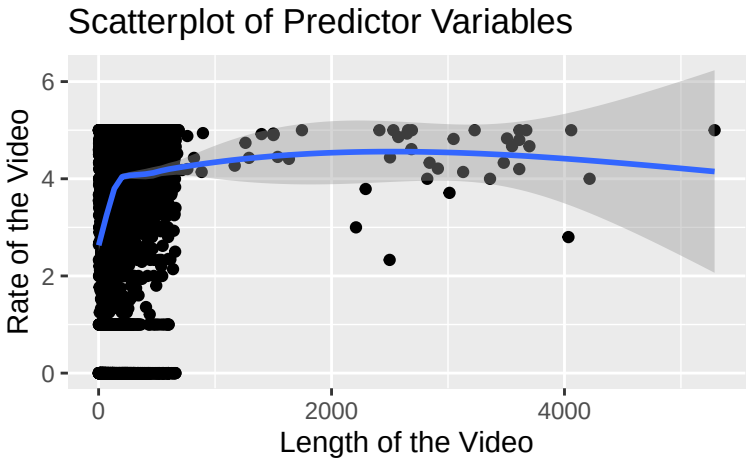

Figure 5: No Evidence of Perfect Colinearity of Predictors

We observe no perfect colinearty between the two predictors. There is not even a hint of near-perfect colinearity. Thus, this assumption is met.

Conclusion: Yes, the assumption is met.

4 Homoskedastic Errors Assumption

For this assumption to be met we expect the variance of the residuals to be constant across the range of predicted values. To test for this, we will look at the scatter plot of model residuals against its fitted values and observe how the variance changes moving left to right.

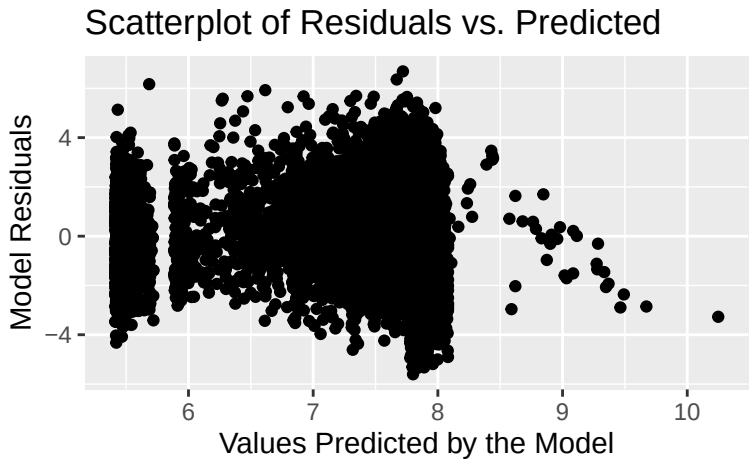

Figure 6: Errors Do Not Appear to be Homoskedastic

In the residual vs. predicted plot, we see what appears to be evidence of heteroskedastic errors. Moving from left to right along the range of predicted values, we clearly see the variance span [-4, 4], then shift to [-2, 4], then grow to [-5, 5], and finish with [-4, 2]. These fluctuations in variance across the range of predicted values is a clear indication of heteroskedasticity.

The Breusch-Pagan test is another way to evaluate for the presence of heteroskedasticity. This test evaluates the null hypothesis: "no evidence for heteroskedastic error variance", and so rejecting the null means we have evidence of a problem.

```
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 128.39, df = 2, p-value < 2.2e-16
```

With a p-value well below 0.05, we reject the null hypothesis. The evidence suggests we cannot rule out the presence of heteroskedasticity.

Conclusion: No, the assumption is *not* met.

5 Normally Distributed Errors Assumption

For this assumption to be met we expect the residuals to be normally distributed. To test for this we look at a histogram of the errors plotted against a reference normal distribution. We also look at the Q-Q plot.

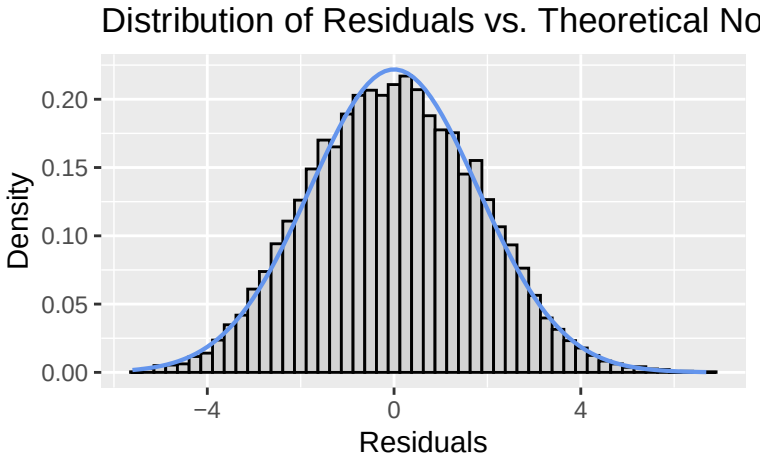

Figure 7: Errors Appear to be Normally Distributed

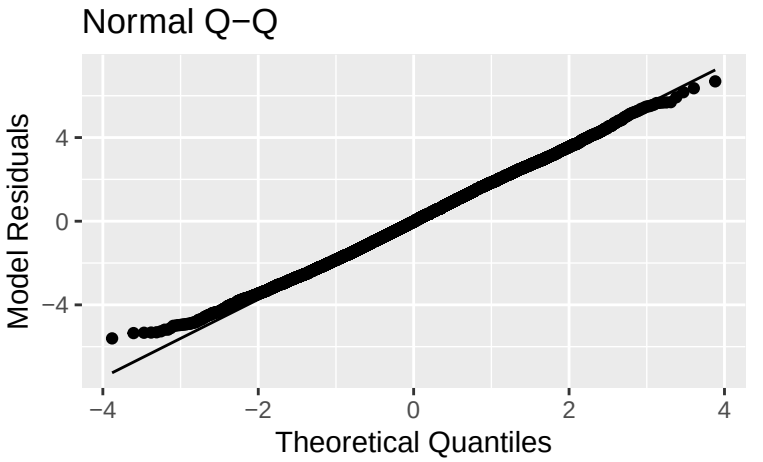

Figure 8: Errors Appear to be Normally Distributed

Neither plots show a concerning amount of deviation from normal. Based on the histogram and Q-Q plot, we see a residual distribution that is pretty faithful to the normal distribution.

Conclusion: Yes, the assumption is met.